Chapter 14

Section 14.1

Check Your Understanding, Page 8:

1. The model is $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. The relationship between percent returning and the number of new birds that join the colony for both species has a slope that is represented by β_1. This represents the predicted change in the number of new birds for an increase of 1 in the percent of birds returning. The *y* intercept for the first species is β_0 and for the second species is $\beta_0 + \beta_2$. These values represent the predicted value of new birds for each species if no birds returned from the previous year.

Check Your Understanding, Page 13:

1. The regression model for the unexposed parents is $\hat{y} = 18.0848 - 5.411$ Humidity and for the exposed parents is $\hat{y} = 18.9332 - 5.411$ Humidity. The slope is –5.411 which means that for each increase in nest humidity index of one unit, the nestling mass is predicted to decrease by 5.411 grams.
2. Because the variable "Exposed" takes on the value 1 for those birds exposed to fleas, and the coefficient for "Exposed" is positive, we predict that the nestling mass will be higher for those birds exposed to fleas during egg laying.
3. If the nest humidity index is 1.2, we predict the nestling mass for non-exposed birds to be $\hat{y} = 18.0848 - 5.411(1.2) + 0.8484(0) = 11.5916$. For exposed birds, we predict a nestling mass of $\hat{y} = 18.0848 - 5.411(1.2) + 0.8484(1) = 12.4400$.
4. Our prediction of the nestling mass will typically be off by about 1.01583 grams from the actual nestling mass.
5. The multiple linear regression model accounts for 47.7% of the variation in nestling mass.

Check Your Understanding, Page 19:

1. Linear: The scatterplot must show a linear pattern for each group of birds and the residual plot should have no leftover patterns. Independent: The measured nestling masses must be independent of each other. Normal: The residuals must appear as if they could have come from a Normal distribution. Equal SD: The residuals should show a similar scatter across the residual plot. Random: The observations need to be randomly selected.
2. From the computer output, $F = 15.51$ and the *P*-value is approximately 0. Because the *P*-value of approximately 0 is less than $\alpha = 0.05$ we reject the null hypothesis. There is convincing evidence that at least one of the coefficients β_1 or β_2 differs from 0. This suggests that at least one of the explanatory variables is useful in predicting nestling mass.
3. The output gives each of the test statistics and *P*-values. For β_0, the test statistic is $t = 27.43$ and the *P*-value is approximately 0. Because the *P*-value is less than $\alpha = 0.05$, we reject the null hypothesis and have convincing evidence that the *y* intercept for the unexposed birds is significantly different from 0. For β_1, the test statistic is $t = -3.93$ and the *P*-value is approximately 0. Because the *P*-value is less than $\alpha = 0.05$, we reject the null hypothesis and have convincing evidence that the slope for both models is significantly different from 0. For β_2, the test statistic is $t = 2.37$ and the *P*-value is 0.024. Because the *P*-value is less than $\alpha = 0.05$, we reject the null hypothesis and have convincing evidence that the difference in the *y* intercept between the exposed birds and the unexposed birds is significantly different from 0.

4. We use a *t* distribution with 34 degrees of freedom. This means that $t^* = 2.032$ so the confidence intervals are: For β_0, $18.0848 \pm 2.032(0.6592) = (16.75, 19.42)$, for β_1, $-5.411 \pm 2.032(1.377) = (-8.21, -2.61)$, and for β_2, $0.8484 \pm 2.032(0.3587) = (0.12, 1.58)$. In all three cases, 0 is not contained in the interval so we conclude that all three coefficients are significantly different from 0. In addition, we also have an interval of plausible values for each of these parameters.

Check Your Understanding, Page 23:

1. The model is $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$. In this case, β_0 is the *y* intercept for the straight line relating the percent of birds returning and the number of new birds that join the colony for the first species of bird and β_1 represents the slope of that line. The other two coefficients refer to the difference between the two bird species. Specifically, β_2 represents the difference in *y* intercepts between the two species and β_3 represents the difference in slopes between the two species.

Check Your Understanding, Page 24:

In the simple linear regression, the typical prediction error is *s* = 20.33460 and the model accounts for 73.2% of the variability in mean SAT score. With the multiple regression model, the typical prediction error is smaller (*s* = 17.4035) and a larger percent of the variability in mean SAT is accounted for (r^2 = 81.2%).

Case Closed, Page 38:

1. Predicted GPA = 0.3267 + 0.14596(HSM) + 0.03591(HSS) + 0.05529(HSE) + 0.0009436(SATM) – 0.0004078(SATCR).

2. Predicted GPA = 0.3267 + 0.14596(10) + 0.03591(10) + 0.05529(10) + 0.0009436(750) – 0.0004078(760) = 3.09607. Residual = 3.86 – 3.10 = 0.76. This student's GPA was 0.76 higher than predicted based on his or her high school math, science, and English grades and SAT math and critical reading scores.

3. STATE: H_0: $\beta_{HSM} = \beta_{HSS} = \beta_{HSE} = \beta_{SATM} = \beta_{SATCR} = 0$ versus H_a: at least one of the β's isn't 0. α = 0.05.
PLAN: The conditions are assumed to be met, so we will proceed with an ANOVA *F* test.
DO: *F* = 11.69, dfN = 5, dfD = 218, *P*-value $\approx$ 0.
CONCLUDE: Because the *P*-value of approximately 0 is less than α = 0.05, we reject H_0. There is convincing evidence that at least one of the β's is not equal to 0.

4. All of the *P*-values are greater than α = 0.05 except for HSM. The *P*-value for high school math grades is approximately 0. This means that we have convincing evidence that high school math grades help predict college GPA, even when all the other variables are included in the model. Because the *P*-values for the other variables are all greater than α = 0.05, we do not have convincing evidence that any of these variables individually help predict college GPA when the other variables are included in the model.

5. You could drop HSS because it has such a high correlation with HSE (*r* = 0.579). If HSE is included in the model, we won't get much unique information from HSS. Likewise, because there is a high correlation between SATCR and SATM (*r* = 0.464), we can drop SATCR as long as we keep SATM in the model.

6. When using the model with HSM, HSE, and SATM to predict GPA, our predictions will typically be off by about 0.698805. Also, 20.7% of the variability in GPA is accounted for by our model using HSM, HSE, and SATM.

7. Linear: There is no obvious pattern leftover in the residual plot. Independent: 224 < 10% of all students at this university; knowing one student's GPA shouldn't help us predict the GPA for another student. Normal: The distribution of residuals doesn't show any strong skewness or outliers. Equal SD: The variability of the residuals looks about the same for all fitted values in the residual plot. Random: The researchers selected a random sample of students from this university.

8. The second model might be preferred because it is much simpler, using only 3 explanatory variables instead of 5. Also, the *R*-sq value is only a little smaller for the simpler model.

Exercises, Page 41:

14.1 Use the model $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ where $y =$ height of the children, $x_1 =$ weight of the children, and $x_2 = 0$ for boys and $x_2 = 1$ for girls. In this model β_0 is the *y* intercept for the line relating weight and height for boys, β_1 is the slope for both lines, and β_2 is the difference in the *y* intercept for boys and girls (girls – boys).

14.2 Let $x_2 = 0$ for the short icicles and $x_2 = 1$ for the longer icicles. At time 0, the short icicles are 10 cm so $\beta_0 = 10$. The long icicles are 20 cm at time 0 so for the long icicles we need a *y* intercept of 20. This means we need to add 10 to the *y* intercept for the short icicles and that means that $\beta_2 = 10$. Finally both sets of icicles have the same growth over time which makes $\beta_1 = 0.15$. The model is $\hat{y} = 10 + 0.15x_1 + 10x_2$.

14.3 (a) For 2003: predicted reporting percent = 89.896 – 0.0643(reporting date). For 2004: predicted reporting percent = 85.967 + 0.0588(reporting date).
(b) In 2003, as the reporting date increases by one, the percent reporting is predicted to decrease by 0.0643%. In 2004, as the reporting date increases by one, the percent reporting is predicted to increase by 0.0588%.
(c) For reporting date 25, we would predict that 89.896 – 0.0643(25) = 88.2885% would report in 2003 and 85.967 + 0.0588(25) = 87.437% would report in 2004. The actual values were 91.2% for 2003 and 90.66% for 2004. This leads to residuals of 91.2% – 88.2885% = 2.9115% for 2003 and 90.66% – 87.437% = 3.223% for 2004.
(d) Yes, we would be willing to use the multiple linear regression model with equal slopes for 2003 and 2004. Despite the fact that the slope in 2003 is negative and the slope for 2004 is positive, they are both close enough to 0 that the slopes could be the same.

14.4 (a) For 1985: predicted percent reporting = 18.943 + 0.2364(reporting period). For 1997: predicted percent reporting = 25.705 + 1.6592(reporting period).
(b) In 1985, as the reporting date increases by one, the percent reporting is predicted to increase by 0.2364%. In 1997, as the reporting date increases by one, the percent reporting is predicted to increase by 1.6592%.
(c) For reporting date 26, we would predict that 18.943 + 0.2364(26) = 25.0894% would report in 1985 and 25.705 + 1.6592(26) = 68.8442% would report in 1997. The actual were reporting percentages 14.2% for 1985 and 62.1% for 1997. This leads to residuals of 14.2% – 25.0894% = –10.8894% for 1985 and 62.1% – 68.8442% = –6.7442 for 1997. It isn't surprising that both residuals are negative, as period 26 is during Christmas time.
(d) No, we would not be willing to use the multiple linear regression model with equal slopes for 1985 and 1997. Both from the graph and from the computer output, it is clear that the slopes are quite different in the two years.

14.5 (a) β_0 gives the *y* intercept for the line relating reporting percentage to reporting date for 2003. β_1 gives the slope of the line relating reporting percentage to reporting date for both years. β_2 gives the difference in *y* intercept between the lines for 2003 and 2004 (2004 – 2003). The estimate for β_0 is 89.049, the estimate for β_1 is 0.0009, and the estimate for β_2 is –2.299.
(b) For 2003, the estimated regression line is $\hat{y} = 89.049 + 0.0009x_1 - 2.299(0) = 89.049 + 0.0009x_1$. For 2004 the estimated regression line is $\hat{y} = 89.049 + 0.0009x_1 - 2.299(1) = 86.75 + 0.0009x_1$.

14.6 For 2000 we get $\hat{y} = 94.9 - 0.717x_1 - 17.8(0) = 94.9 - 0.717x_1$ and for 1998 we get $\hat{y} = 94.9 - 0.717x_1 - 17.8(1) = 77.1 - 0.717x_1$.
(b) The lines we computed in part (a) are identical to the lines computed in the example.
(c) No, the regression standard error will not change because we have not added (or eliminated) any variables. We just redefined one variable.

14.7 (a) The output gives $a = 1.28071$ and because *a* estimates $\log(\alpha)$, we get $1.28071 = \log(\alpha)$ or $10^{1.28071} = 19.09 = \alpha$. And because *b* estimates β and the output gives $b = 0.82179$, our estimate of β is 0.82179.
(b) The model from the output is $\widehat{\log}(MR) = 1.28071 + 0.82179\log(BM)$ so for a worm with body mass of 3 g, we predict $\widehat{\log}(MR) = 1.28071 + 0.82179\log(3) = 1.6728$. This means that we predict a metabolic rate of $10^{1.6728} = 47.08$ μl/min.
(c) The simple linear regression relating log metabolic rate to log body mass for tobacco hornworm caterpillars accounts for 93.7% of the variability in the log metabolic rate of these caterpillars.

14.8 (a) Linear: The residual plot possibly shows a leftover M-shaped pattern which might suggest that a different model would be better. If all other conditions are met, we should proceed with caution. Independent: Because the caterpillars were selected at random, the metabolic rates of one caterpillar should be independent of the metabolic rates of another caterpillar. Also, the sample can be assumed to be less than 10% of the population of all caterpillars. Normal: The histogram shows us that there is no strong skewness or outliers. Also, with the large sample size, we can proceed as if the residuals have a Normal distribution. Equal SD: The residual plot shows us that the standard deviation of *y* is approximately the same for each value of *x*. Random: The sample of caterpillars was selected at random.
(b) The were 209 caterpillars in the study so the degrees of freedom are $209 - 2 = 207$. This gives $t^* = 1.97$ and a 95% confidence interval of $0.82179 \pm 1.97(0.01477) = (0.79269, 0.85089)$.
(c) Neither $\frac{2}{3} = 0.667$ nor $\frac{3}{4} = 0.75$ are in the interval, so neither one is a plausible for β in this setting.

14.9 (a) It is reasonable to use a multiple linear regression model with parallel lines in this case because it appears that as the length of service increases, the wages increase at about the same rate for women in large and small banks.
(b) β_0 represents the average salary of women at small banks who are new employees (LOS = 0). β_1 represents the average change in wages for each additional month of service for women at both small and large banks. β_2 represents the difference in starting salaries (LOS = 0) for women at large banks and women at small banks (large – small). In this case our estimate of β_0 is 37.565, our estimate of β_1 is 0.08289, and our estimate of β_2 is 8.916.
(c) The least-squares line for small banks is $\hat{y} = 37.565 + 0.08289x_1 + 8.916(0) = 37.565 + 0.08289x_1$. The least-squares line for large banks is $\hat{y} = 37.565 + 0.08289x_1 + 8.916(1) = 46.481 + 0.08289x_1$.
(d) Using the regression equations, we predict wages of $\hat{y} = 37.565 + 0.08289(120) = 47.51$ for a woman at a small bank, and wages of $\hat{y} = 46.481 + 0.08289(120) = 56.43$ for a woman at a large bank.

14.10 (a) Looking back at the scatterplot from before exercises 7 and 8, it appears that it would be reasonable to use a multiple linear regression model with parallel lines because it appears that the relationship between body mass and metabolic rate is the same for both stages of caterpillars.
(b) β_0 represents the *y* intercept for the line relating log metabolic rate to log body mass for Stage 4 caterpillars. β_1 represents the average change in log metabolic rate for each additional log gram of weight for caterpillars at both Stage 4 and Stage 5. β_2 represents the difference in *y* intercept for caterpillars in Stage 5 from those in Stage 4 (stage 5 – stage 4). In this case our estimate of β_0 is 1.23917, our estimate of β_1 is 0.69828, and our estimate of β_2 is 0.1468.
(c) The least-squares line for Stage 4 caterpillars is $\hat{y} = 1.23917 + 0.69828x_1 + 0.1468(0)$ $= 1.23917 + 0.69828x_1$. The least-squares line for Stage 5 caterpillars is $\hat{y} = 1.23917 + 0.69828x_1 + 0.1468(1) = 1.38597 + 0.69828x_1$.
(d) Using the regression equation, we find a value of $\hat{y} = 1.23917 + 0.69828(\log 3) = 1.57233$ for a Stage 4 caterpillar. This translates into a metabolic rate of $10^{1.57233} = 37.35$ μl/min. For a Stage 5 caterpillar the predicted value of *y* is $\hat{y} = 1.38597 + 0.69828(\log 3) = 1.71913$ which gives a metabolic rate of $10^{1.71913} = 52.38$ μl/min.

14.11 The value of R^2 is 29.1%. This means that the multiple linear regression model accounts for 29.1% of the variability in the wages of female employees at small and large banks in Indiana. The output also gives $s = 9.27$ which means that we will typically be off by 9.27 units when predicting wages. Note that the original values were rescaled for confidentiality, so the standard deviation of the residuals is also rescaled by the same amount.

14.12 The value of R^2 is 94.5%. This means that the multiple linear regression model accounts for 94.5% of the variability in the log metabolic rate for tobacco hornworm caterpillars in Stages 4 or 5. The output also gives $s = 0.1$ which means that we will typically be off by about 0.1 when predicting the log of metabolic rate.

14.13 (a) Linear: The residual plot does not show any leftover pattern that would suggest a different model would be better. Independent: Because the women were selected at random, the wages of one woman should be independent of the wages of another woman. Also, *n* = 59 < 10% of all female bank tellers in Indiana. Normal: The histogram of residuals shows no strong skewness or outliers. Equal SD: The residual plot shows us that the standard deviation of *y* is approximately the same for each *x*. Random: The sample of women was selected at random.
(b) The output in Exercise 9 gives $F = 11.5$ with a *P*-value of approximately 0. This tests the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ against the alternative that at least one of the β's is not 0. Because the *P*-value of approximately 0 is less than α = 0.05, we reject the null hypothesis. We have convincing evidence that the two explanatory variables together help to predict the wages of women who work in banks in Indiana.

14.14 (a) Linear: The residual plot does show a leftover M-shaped pattern which might suggest that a different model would be better. If all other conditions are met, we should proceed with caution. Independent: Because the caterpillars were selected at random, the metabolic rate of one caterpillar should be independent of the metabolic rate of another caterpillar. Also, the sample size is less than 10% of all caterpillars. Normal: The histogram shows no strong skewness or outliers. Equal SD: The residual versus fitted values plot shows us that the standard deviation of *y* is approximately the same for each *x*. Random: The sample of caterpillars was selected at random.

(b) The output in exercise 10 gives $F = 1784.89$ with a *P*-value of approximately 0. This tests the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ against the alternative that at least one of the β's is not 0. Because the *P*-value of approximately 0 is less than α = 0.05, we reject the null hypothesis. We have convincing evidence that the two explanatory variables together help to predict the log(metabolic rate) of caterpillars.

14.15 (a) Going back to the output in Exercise 9, we see that the *P*-values for all three tests are very small. The *P*-value for the test for β_0 is approximately 0 and the other two are listed as being 0.001. Because each of the *P*-values is small, we would reject the null hypothesis that the parameter is 0 in all three cases. This means that all three parameters are important to the model.
(b) Using df = 56, the appropriate critical value for a 95% confidence interval is $t^* = 2.00$. The resulting interval is $0.0829 \pm 2.00(0.02349) = (0.036, 0.130)$. We are 95% confident that the interval from 0.036 to 0.130 captures the true slope of the regression line relating wages to length of service.

14.16 (a) Going back to the output in Exercise 10, we see that the *P*-values for all three tests are very small. In fact, they are all listed as being approximately 0. Because each of the *P*-values is small, we would reject the null hypothesis that the parameter is 0 in all three cases. This means that all three parameters are important to the model.
(b) Using df = 206, the appropriate critical value for a 95% confidence interval is $t^* = 1.97$. The resulting interval is of $0.69828 \pm 1.97(0.02628) = (0.647, 0.750)$. We are 95% confident that the interval from 0.647 to 0.750 captures the true slope of the regression line relating log metabolic rate to log body mass.

14.17 The regression model is $\hat{y} = 0 + 20x_1 + 8x_1x_2$ where $x_2 = 0$ for buses and $x_2 = 1$ for self-guided cars. This gives the models $\hat{y} = 20x_1 + 8x_1(0) = 20x_1$ for buses and $\hat{y} = 20x_1 + 8x_1(1) = 20x_1 + 8x_1 = 28x_1$ for self-guided cars.

14.18 The regression model is $\hat{y} = 10 + 0.15x_1 + 10x_2 - 0.03x_1x_2$ where $x_2 = 0$ for the 10 cm icicles and $x_2 = 1$ for the 20 centimeter icicles. This gives the models
$\hat{y} = 10 + 0.15x_1 + 10(0) - 0.03x_1(0) = 10 + 0.15x_1$ for 10 cm icicles and
$\hat{y} = 10 + 0.15x_1 + 10(1) - 0.03x_1(1) = 10 + 0.15x_1 + 10 - 0.03x_1 = 20 + 0.12x_1$ for the 20 cm icicles.

14.19 Use the model $\mu_y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2$ where $x_2 = 0$ for boys and $x_2 = 1$ for girls. In this case β_0 represents the *y* intercept of the line for boys, β_1 represents the slope of the line for boys, β_2 represents the difference in the *y* intercepts of the lines for boys and girls (girls – boys), and β_3 represents the difference in the slopes of the lines for boys and girls (girls – boys).

14.20 (a) Yes, it is likely that waist size and body fat are positively correlated. That is, a man who has a large amount of body fat probably also has a large waist, whereas a man who has a small amount of body fat probably has a small waist.
(b) If you take two men who have the same waist size, but one of them is taller, the taller man is likely to be proportionally thinner and therefore have a smaller percentage of body fat.
(c) Yes, it is likely that the parameter estimate for height would become negative if we also use waist size in the model. In that case, the parameter measures the relationship between height and body fat for a given waist size. As discussed in part (b), this relationship is likely to be negative.

14.21 (a) For states with 40% or more taking the SAT, the estimated regression line is $\hat{y} = 516.71 - 0.2488x_1 + 78.51(0) - 2.5102x_1(0) = 516.71 - 0.2488x_1$. For the other states, the estimated regression line is $\hat{y} = 516.71 - 0.2488x_1 + 78.51(1) - 2.5102x_1(1) = 595.22 - 2.759x_1$.

(b) Use the second regression line from part (a) to get $\hat{y} = 595.22 - 2.759(25) = 526.245$.

(c) The output gives $t = -5.65$ with a *P*-value of approximately 0. Because the *P*-value of approximately 0 is less than $\alpha = 0.05$, we reject the null hypothesis that $\beta_3 = 0$. In other words, the interaction term is important and we should not fit the more restrictive model with parallel lines.

(d) The multiple linear regression model accounts for 88.4% of the variation in the state mean SAT critical reading scores. When using this model, our predictions will typically be off by about 13.8 points.

14.22 (a) For female subjects, the estimated regression line is $\hat{y} = 201.2 + 24.026x_1 + 509.3(0) - 7.275x_1(0) = 201.2 + 24.026x_1$. For male subjects, the estimated regression line is $\hat{y} = 201.2 + 24.026x_1 + 509.3(1) - 7.275x_1(1) = 710.5 + 16.751x_1$.

(b) Use the regression line for females: $\hat{y} = 201.2 + 24.026(42.0) = 1210.29$ cal/24 hours.

(c) The output gives $t = -0.78$ with a *P*-value of 0.447. Because the *P*-value of 0.447 is greater than $\alpha = 0.05$, we fail to reject the null hypothesis that $\beta_3 = 0$. In other words, we don't have convincing evidence that the interaction term is important, so we can fit the more restrictive model with parallel lines.

(d) The multiple linear regression model accounts for 80.7% of the variation in the metabolic rate. When using this model, our predictions will typically be off by about 123.815 cal/24hours.

14.23 (a) No, the plots do not indicate any problems with the conditions for inference about the regression model. The residual plot shows no obvious leftover pattern and the standard deviation of *y* is about the same for each value of *x*. The histogram of residuals is reasonably symmetric and has no outliers.

(b) From the output in Exercise 21, $F = 117.24$ and the *P*-value is approximately 0. Because the *P*-value of approximately 0 is less than $\alpha = 0.05$, we reject the null hypothesis that $\beta_1 = \beta_2 = \beta_3 = 0$. We have convincing evidence that at least one of the explanatory variables is useful in predicting mean SAT critical reading scores.

(c) The *P*-value for β_1 is 0.269. Because this is greater than $\alpha = 0.05$, we do not have enough evidence to conclude that β_1 is not 0. This suggests that the slope for those states in which more than 40% of high school graduates taking the SAT could be 0. All other *P*-values are approximately 0. We have convincing evidence that there is a slope that is different from 0 for the states where fewer than 40% take the test, that the *y* intercept is different from 0 for states where more than 40% take the test, and that the *y* intercept is different for the two groups of states.

14.24 (a) No, the plots do not indicate any major problems with the conditions for inference about the regression model. The residual plot shows no leftover pattern that would indicate that the model is incorrect and the standard deviation of *y* is about the same for each value of *x*. The histogram is somewhat skewed to the right but not dramatically so and has no outliers.

(b) From the output in Exercise 22, $F = 20.95$ and the *P*-value is approximately 0. Because the *P*-value of approximately 0 is less than $\alpha = 0.05$, we reject the null hypothesis that $\beta_1 = \beta_2 = \beta_3 = 0$. We have convincing evidence that at least one of the explanatory variables is useful in predicting metabolic rate.

(c) The only *P*-value that is less than $\alpha = 0.05$ is the one for mass. We have convincing evidence that mass is useful for predicting metabolism. Because the coefficient for the interaction term could be 0, this means that a model with parallel lines could be appropriate (we wouldn't need different slopes for the two genders). Because the coefficient for gender could be 0, and since we've determined that parallel lines

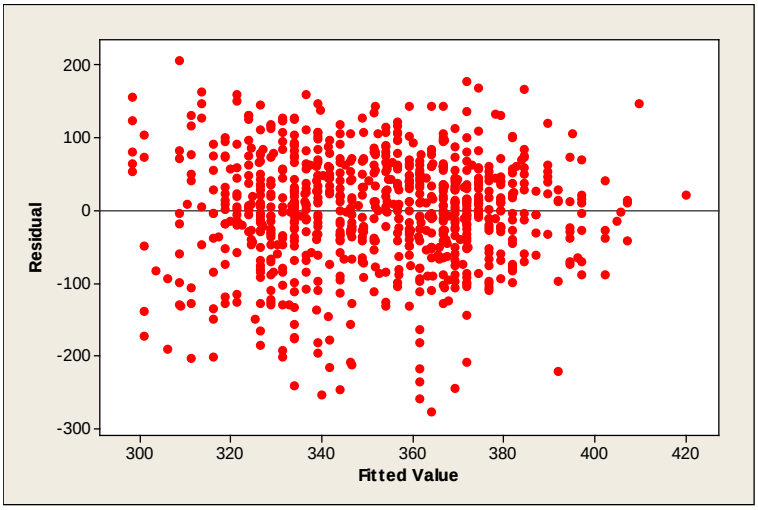
are ok, this means only one *y* intercept is needed. When we put all of this together, we find that we can use a single line with a *y* intercept of 0.

14.25 (a) The Minitab output for this model is given below. It shows that the estimated regression model is $\hat{y} = 321.116 + 2.5373Hgt90 - 37.802Fert$. The standard deviation of the residuals is $s = 76.5506$ and $R^2 = 8.3\%$. While all of the coefficients are found to be something other than 0 (all *P*-values are very small), the typical error in predicting the height of the trees is 76 cm, which is a lot, given that the trees are only between 300 and 400 cm tall in 1997. Also, we are only accounting for only 8.3% of the variability in tree height by using our model. Finally, the graph of the residuals versus fitted values (predicted values) shows that the standard deviation of *y* is about the same for each *x* and does not show any patterns to be concerned about.

```
Predictor      Coef  SE Coef      T      P
Constant    321.116    9.759  32.90  0.000
Hgt90        2.5373   0.4790   5.30  0.000
Fert        -37.802    5.403  -7.00  0.000

S = 76.5506   R-Sq = 8.3%   R-Sq(adj) = 8.0%

Analysis of Variance
Source           DF       SS      MS      F      P
Regression        2   425196  212598  36.28  0.000
Residual Error  806  4723158    5860
Total           808  5148354
```



(b) We would expect that the height of the trees in 1996 would be highly correlated with their height in 1997 because we are measuring the same thing in time periods that are relatively close together. The Minitab output for this new model is given below. This model is much better. The standard deviation of the residuals has been reduced to 18.63 and R^2 has increased to 94.4%.

```
Predictor      Coef  SE Coef       T      P
Constant     44.964    3.470   12.96  0.000
Hgt90       -0.1590   0.1195   -1.33  0.183
Fert         -2.390    1.361   -1.76  0.079
Hgt96       1.09409  0.00992  110.24  0.000

S = 18.6347   R-Sq = 94.4%   R-Sq(adj) = 94.3%
```

```
Analysis of Variance
Source           DF       SS       MS        F      P
Regression        3  4634006  1544669  4448.28  0.000
Residual Error  796   276411      347
Total           799  4910418
```

(c) The *t* statistic for *Hgt90* went from being positive to negative and from being significant to being non-significant in moving from the model in part (a) to the model in part (b). Basically, the information in *Hgt90* does not help us predict height in 1997 if we can use *Hgt96*.
(d) The Minitab output is given below. In this model, all explanatory variables are significant because their *P*-values are all less than $\alpha = 0.05$. This means that each of the explanatory variables contributes additional useful information, even when the other variables are factored in. Also, the standard deviation of the residuals is the smallest of the three models explored, and R^2 is the largest of the models explored. This is the best of the three models.

```
Predictor     Coef  SE Coef      T      P
Constant    45.680    2.876  15.88  0.000
Diam97      2.1432   0.7203   2.98  0.003
Fert        -2.678    1.297  -2.07  0.039
Hgt96      1.03756  0.01909  54.36  0.000

S = 18.2935   R-Sq = 94.8%   R-Sq(adj) = 94.8%

Analysis of Variance
Source           DF       SS       MS        F      P
Regression        3  5127726  1709242  5107.51  0.000
Residual Error  845   282782      335
Total           848  5410508
```

14.26 (a) The Minitab output is given below. Because the *P*-value for the indicator variable is greater than $\alpha = 0.05$, we do not have convincing evidence that the relationship between measured and self-estimated reading ability is different for both boys and girls.

```
Predictor    Coef  SE Coef     T      P
Constant   24.381    8.084  3.02  0.004
EST        10.599    2.131  4.97  0.000
IndGender   6.708    5.178  1.30  0.200

S = 19.8432   R-Sq = 32.3%   R-Sq(adj) = 30.0%

Analysis of Variance
Source           DF       SS      MS      F      P
Regression        2  10720.3  5360.1  13.61  0.000
Residual Error   57  22443.9   393.8
Total            59  33164.2
```

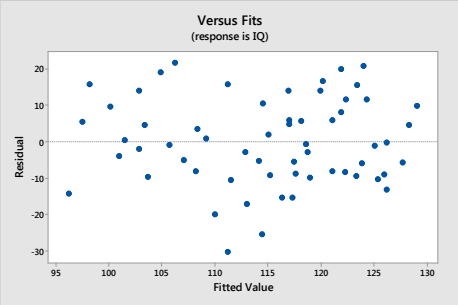
(b) The Minitab output is given below. While the model is significant (the *P*-value for the *F* test is approximately 0), the standard deviation of the residuals is fairly high (12.24 for values that are around 100) and the R^2 is only moderate at 35.1%.

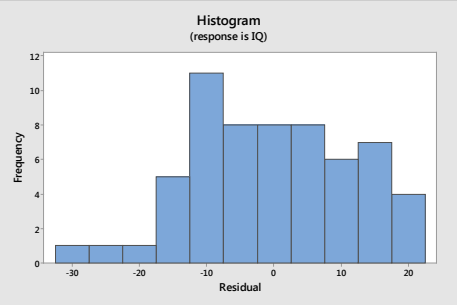
```
Predictor     Coef  SE Coef      T      P
Constant    85.179    7.908  10.77  0.000
LSS          0.513    1.390   0.37  0.713
READ       0.22869  0.08181   2.80  0.007
EST          3.472    1.606   2.16  0.035

S = 12.2420   R-Sq = 35.1%   R-Sq(adj) = 31.6%
```

```
Analysis of Variance
Source          DF       SS      MS      F      P
Regression       3   4532.5  1510.8  10.08  0.000
Residual Error  56   8392.5   149.9
Total           59  12925.0
```

(c) The residuals versus fits graph shows that the standard deviation of *y* is basically the same for each *x* and there are no leftover patterns to be concerned about. The histogram of the residuals doesn't show strong skewness or outliers. It appears that our conditions have been met.





(d) The Minitab output is given below. None of the explanatory variables are now significant. This is because we are testing whether an individual variable adds more information *after* all of the other variables have been entered into the equation. In this case, both variables and their interaction carry essentially the same information so there is nothing to be gained from, say READ, after we have already used the information from EST and Read*EST.

```
Predictor     Coef  SE Coef     T      P
Constant     87.97    13.31  6.61  0.000
READ        0.2225   0.2394  0.93  0.357
EST          3.402    4.060  0.84  0.406
Read*EST   0.00327  0.06394  0.05  0.959

S = 12.2566   R-Sq = 34.9%   R-Sq(adj) = 31.4%

Analysis of Variance
Source          DF       SS      MS      F      P
Regression       3   4512.4  1504.1  10.01  0.000
Residual Error  56   8412.6   150.2
Total           59  12925.0
```

14.27 c

14.28 b

14.29 c

14.30 b

14.31 c

14.32 (a) The model that is fitted is $\mu_y = \beta_0 + \beta_1 Mph + \beta_2 IndSlow + \beta_3 NoIncline + \beta_4 2\%Incline$. The estimates for the parameters are 64.75 for β_0, 145.841 for β_1, –50.01 for β_2, –145.06 for β_3, –72.83 for β_4, and 34.2865 for σ.

(b) There are 6 different fitted lines and they all have the same slope because no interaction terms were used in the model.

For fast speeds and 4% incline the estimated line is

$\hat{y} = 64.75 + 145.841 Mph - 50.01(0) - 145.06(0) - 72.83(0) = 64.75 + 145.841 Mph.$

For slow speeds and 4% incline the estimated line is

$\hat{y} = 64.75 + 145.841 Mph - 50.01(1) - 145.06(0) - 72.83(0) = 14.74 + 145.841 Mph.$

For fast speeds and 2% incline the estimated line is

$\hat{y} = 64.75 + 145.841 Mph - 50.01(0) - 145.06(0) - 72.83(1) = -8.08 + 145.841 Mph.$

For slow speeds and 2% incline the estimated line is

$\hat{y} = 64.75 + 145.841 Mph - 50.01(1) - 145.06(0) - 72.83(1) = -58.09 + 145.841 Mph.$

For fast speeds and no incline the estimated line is

$\hat{y} = 64.75 + 145.841 Mph - 50.01(0) - 145.06(1) - 72.83(0) = -80.31 + 145.841 Mph.$

For slow speeds and no incline the estimated line is

$\hat{y} = 64.75 + 145.841 Mph - 50.01(1) - 145.06(1) - 72.83(0) = -130.32 + 145.841 Mph.$

(c) The model appears to give a reasonably good fit to the data. R^2 is quite large at 99.3% (meaning we are accounting for 99.3% of the variation in the number of calories burned) and the standard deviation of the residuals is reasonably small at 34.29 calories per hour.

(d) If we want to test whether more calories are burned for higher speeds, we are looking at the parameter of the *Mph* variable. The hypotheses are $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 > 0$. According to the output, the test statistic is 56.17 and the *P*-value is approximately 0 (even after dividing the provided output by 2 to account for the one-sided test). Because the *P*-value is less than $\alpha = 0.05$, we reject the null hypothesis. We have convincing evidence that more calories are burned for higher speeds.